Weekly Report

Yuxin Ma

10.31.2016 - 11.06.2016

Projects

Deep Learning for Visualization

- **Datasets** This week the images of the R datasets[1] are generated, including 700 datasets and 62762 scatterplot images with the size of 100px×100px. The data points are represented as 2-pixel-width blue (RGB(0, 0, 255)) dots with the opacity of 0.4. The Python Matplotlib library is used for generating the images.

  There are several examples of the images listed in Figure 1. The reason to choose a variety of datasets is to cover as many real-world scatterplot patterns as possible, then we can "teach" the model to recognize diversified scatterplot patterns.

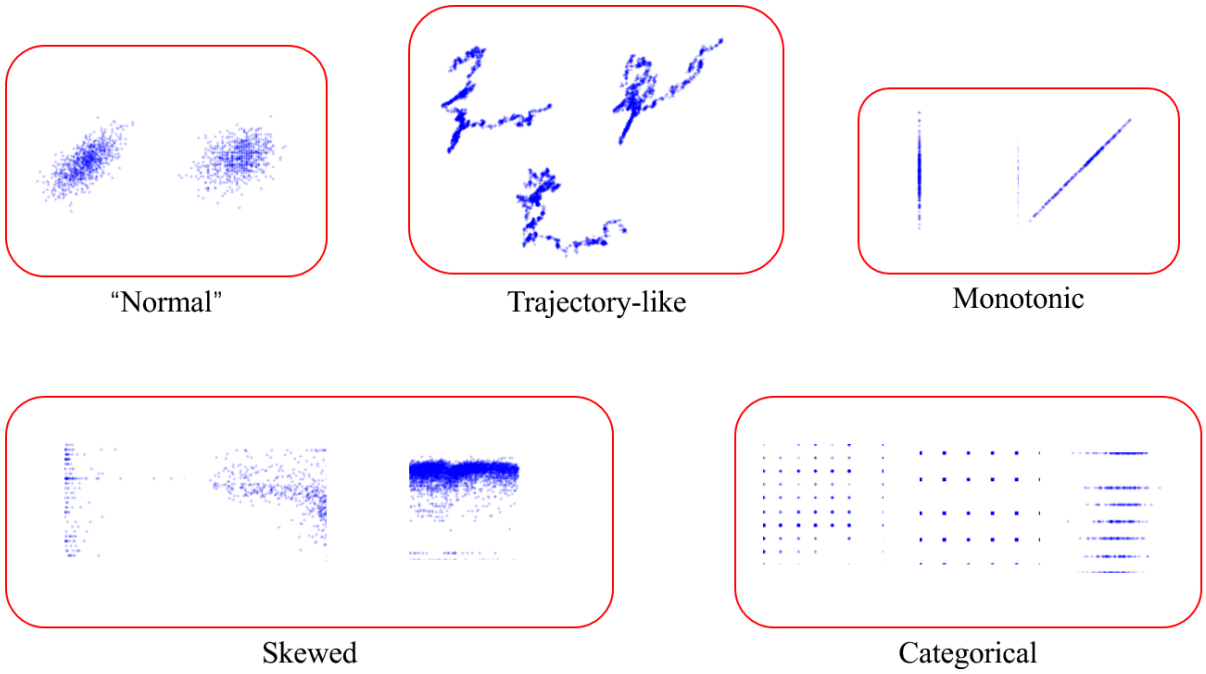

Figure 1: Some examples.

  The 9-dimensional scagnostics values are computed as well. It can be used for using clustering algorithms to get a preliminary clustering results. However after a skip of the images and their scagnostics values I found that the scagnostics values are not such describable for the visual patterns in scatterplots.

- **Method** Initially I have adopted deep convolution auto-encoder (CNN auto-encoder) as the unsupervised feature extraction method. The 100×100 (namely 10000 dimensional by pixel) images are transformed into 1352 dimensional features space by the auto-encoder. The features are then projected with LargeVis [1] on 2-d space. Figure 2 shows the projection result where each point represents a single scatterplot image. It is obvious that many small groups can be discovered.

[1] http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

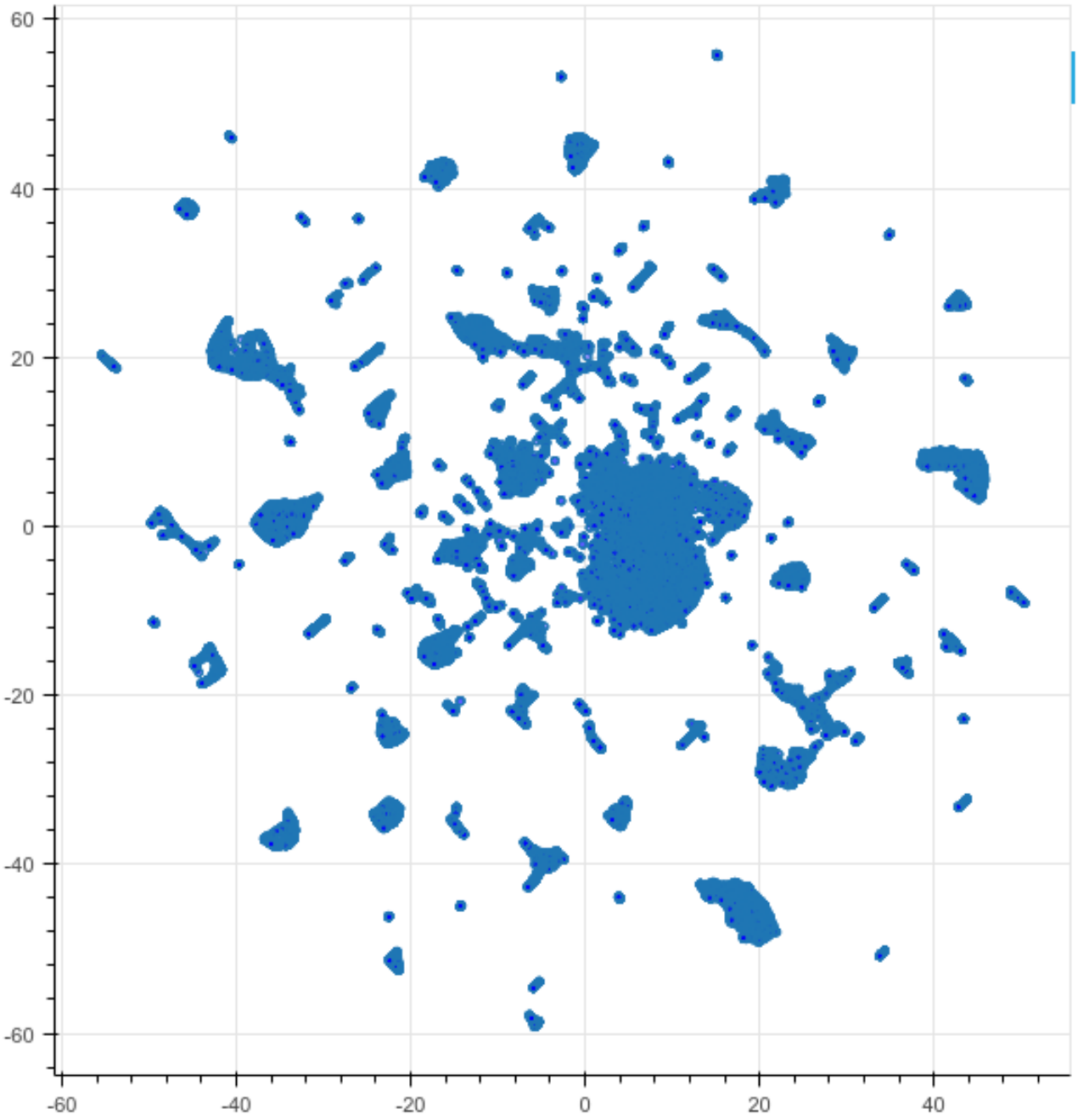

Figure 2: The LargeVis projection result.

In the small groups the distribution of images does not show useful patterns. The original idea is that the scatterplot images in a group should share some kind of similarity, however it is not obvious in the current result. Two sample results are shown below (Figure 34). In the next week I will focus on improving the result.

Plan for the Next Week

- About how to improve the method:
  - **Data**: The scatterplot images contains too many categorical ones which present dot-array-like shape. The dataset will be carefully and manually reviewed in this week.
  - **Method**: The network structure and the parameters will be investigated. Maybe in the next step some supervised model (CNN) can start to be considered.
- Think about applications and tasks.

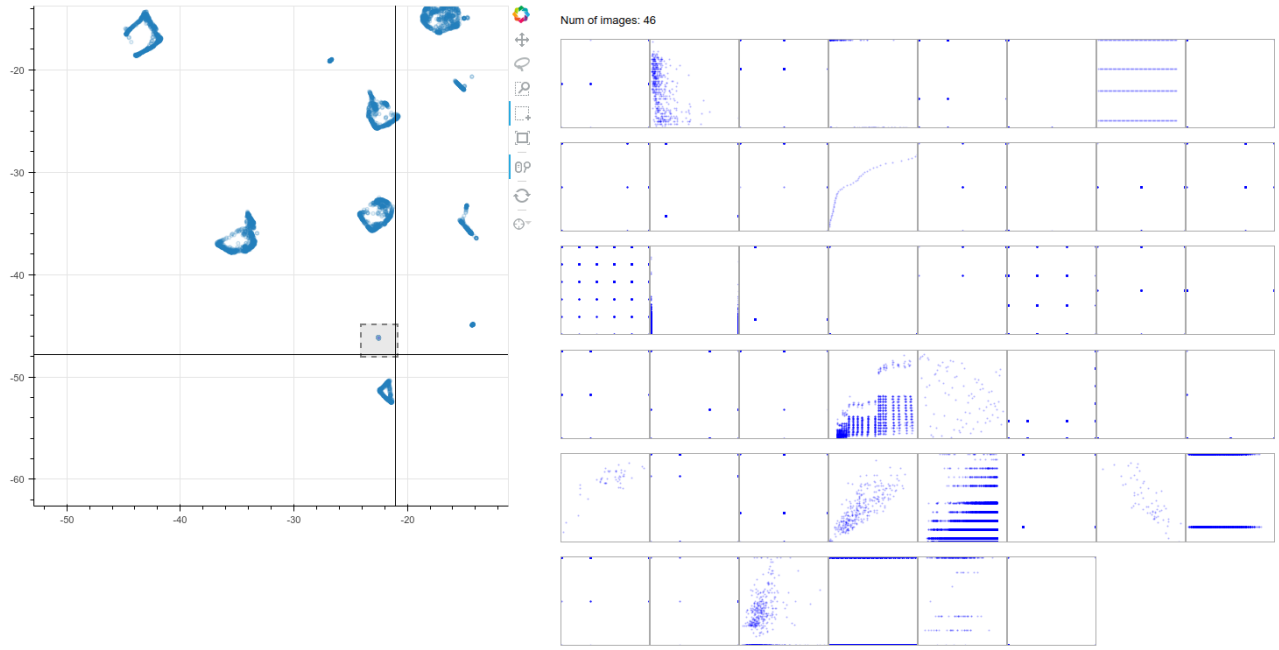

Figure 3: Images in a selected area in the projection result.

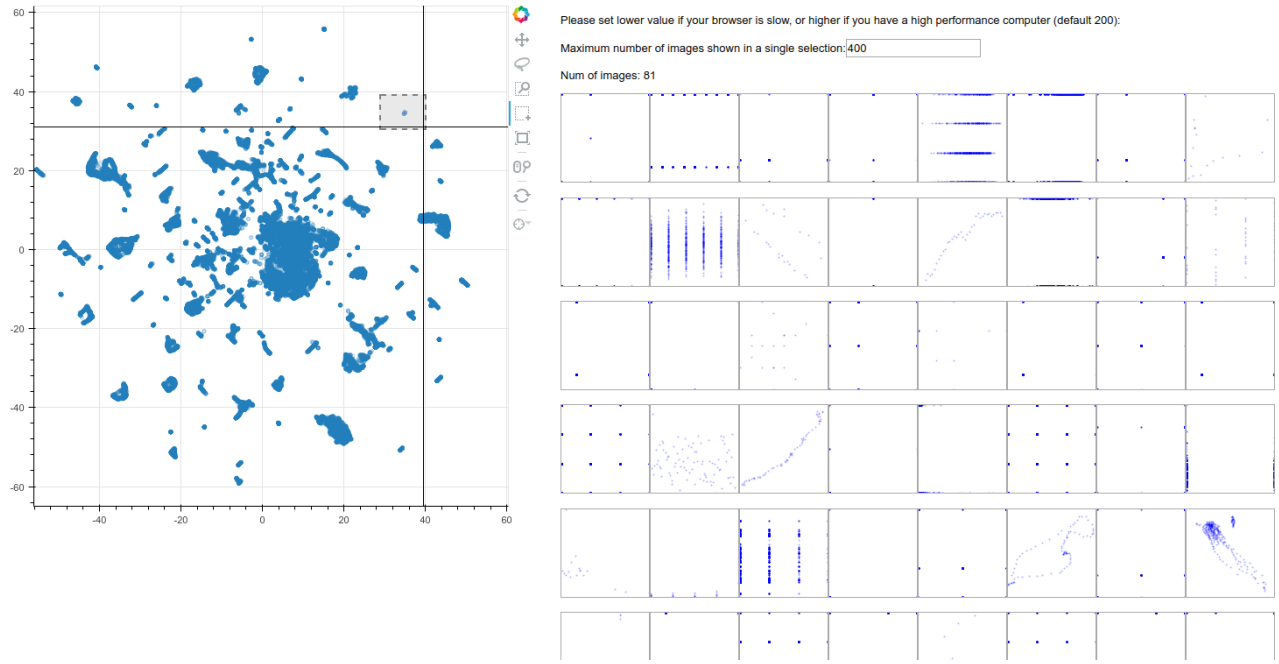

Figure 4: (Continue) Images in a selected area in the projection result.

References

[1] J. Tang, J. Liu, M. Zhang, and Q. Mei, "Visualizing large-scale and high-dimensional data," in *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297, International World Wide Web Conferences Steering Committee, 2016.